Great Learning
POWER AHEAD

SLC Walkthrough for Practice Set Week2

□ Import required libraries.
□ Load data set.

Let's begin with some hands-on practice exercises:

1. Identify and remove the variables in the data which are insignificant for the classification analysis.

HINT:

- Look at the variance or standard deviation for each variable and decide which are the variable are significant or insignificant.
- If incase any id kind of features are present in the dataframe, drop them.

2. Is there any record that is recorded more than once? If yes, do the needful.

HINT:

- Check the duplicates in the dataframe.
- Drop them if there are any duplicates in the data.

3. Are there any outliers present in the data? If yes, remove such observations using the quartiles of the variables.

HINT:

- Check the skewness or distribution plot to know if any outliers are present or not?
- If Yes Then Use IQR method to remove the certain level of outliers.
  Use following formula:
  Q1 = df.quantile(0.25), Q3 = df.quantile(0.75), IQR = Q3 - Q1

  df[~((df < (Q1 - 1.5 * IQR)) | (df> (Q3 + 1.5 * IQR))).any(axis=1)]

4. Is scaling required for the KNN algorithm? If yes, scale the data such that the range of each variable will be between 0 and 1.

HINT:

- Is scaling required for KNN algorithm:
  Hint: KNN is distance-based algorithm.
- If yes: then use scaling such a way that values between 0 to 1.
  Hint: two scaling techniques are standard and min max scaling.

5. Build a 7-NN model on 70% of the data using the 'Chebyshev' distance and find its accuracy.

HINT:

- Split the data into train and test such that sets proportion should be 70:30 □ Build a KNN classifier model with following params.
  - □ Number of neighbors are 7.
  - □ Metrics is Chebyshev.
- Find the accuracy for test data. Use accuracy score method.

6. Find the best value of 'K' for the KNN model from the given list of values (use 5fold cross validation).

HINT:

- Use Grid search CV to do the hyperparameter tuning with five-fold crossvalidation.
- Use best params option from fitted Grid model to get the best param.

7. Build a naive bayes model on 70% of the original data and plot the ROC curve along with the AUC score.

HINT:

- As per the naïve Bayes algorithm, feature scaling is not required,

  Therefore using original data to fit the model.

- Split the original data into train and test such that proportion would be 70:30.
- Build any Naïve Bayes model that's suites for the data.
- Plot ROC AUC curve.

8. Which distance metric among 'manhattan', 'euclidean' and 'chebyshev' is suitable for the given dataset? (consider K = 19).

HINT:

- Use Hyperparameter tuning using Grid Search CV to find the best params.
- Get the best param from the Fitted Grid model using best_params_ option.

9. Find the euclidean distance between the first observation of the dataframe 'X_test_scaled' and its five neighboring points in the train set (use for loop).

HINT:

- Use nested for loop to calculate the Euclidean distance.
- Use the Euclidean distance formula for one test record and 5 train record.
- Append the distance to list.
- Sort and print distance.

10. Use the parameters obtained in Q8 to build the KNN model, and find the number of false predictions using the test set.

HINT:

- Use params that we are selected from grid search cv model.
- Use confusion matrix to know number of flase prediction.